REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	New Reprint	-

4. TITLE AND SUBTITLE

Re-analysis of metagenomic sequences from acute flaccidmyelitis patients reveals alternatives to enterovirus D68 infection

5a. CONTRACT NUMBER	W911NF-14-1-0490
5b. GRANT NUMBER	
5c. PROGRAM ELEMENT NUMBER	611103
5d. PROJECT NUMBER	
5e. TASK NUMBER	
5f. WORK UNIT NUMBER	

6. AUTHORS

Florian P. Breitwieser, Carlos A. Pardo, Steven L. Salzberg

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Johns Hopkins University
Physics & Astronomy
3400 North Charles Street
Baltimore, MD 21218 -2685

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
65316-MA-MUR.2

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

Metagenomic sequence data can be used to detect the presence of infectious viruses and bacteria, but normal microbial flora make this process challenging. We re-analyzed metagenomic RNA sequence data collected during a recent outbreak of acute flaccid myelitis (AFM), caused in some cases by infection with enterovirus D68. We found that among the patients whose symptoms were previously attributed to enterovirus D68, one patient had clear evidence of infection with Haemophilus influenzae, and a second patient had a severe Staphylococcus aureus infection caused by a methicillin-resistant strain. Neither of these bacteria were identified in the original study

15. SUBJECT TERMS

Bioinformatics, Genomics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			David Karig
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 240-228-4719

Report Title

Re-analysis of metagenomic sequences from acute flaccidmyelitis patients reveals alternatives to enterovirus D68 infection

ABSTRACT

Metagenomic sequence data can be used to detect the presence of infectious viruses and bacteria, but normal microbial flora make this process challenging. We re-analyzed metagenomic RNA sequence data collected during a recent outbreak of acute flaccid myelitis (AFM), caused in some cases by infection with enterovirus D68. We found that among the patients whose symptoms were previously attributed to enterovirus D68, one patient had clear evidence of infection with Haemophilus influenzae, and a second patient had a severe Staphylococcus aureus infection caused by a methicillin-resistant strain. Neither of these bacteria were identified in the original study. These observations may have relevance in cases that present with flaccid paralysis because bacterial infections, co-infections or post-infection immune responses may trigger pathogenic processes that may present as poliomyelitis-like syndromes and may mimic AFM. A separate finding was that large numbers of human sequences were present in each of the publicly released samples, although the original study reported that human sequences had been removed before deposition.

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

ARO Report Number 65316.2-MA-MUR
Re-analysis of metagenomic sequences from ac...

Block 13: Supplementary Note

Approved for public release; distribution is unlimited.

RESEARCH ARTICLE

REVISED Re-analysis of metagenomic sequences from acute flaccid myelitis patients reveals alternatives to enterovirus D68 infection [v2; ref status: approved 1, http://f1000r.es/5mz]

Florian P. Breitwieser[1], Carlos A. Pardo[2], Steven L. Salzberg[1,3]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, 21205, USA
[2]Department of Neurology, Johns Hopkins Hospital, Baltimore, MD, 21205, USA
[3]Departments of Biomedical Engineering, Computer Science, and Biostatistics, Johns Hopkins University, Baltimore, MD, 21218, USA

v2 **First published:** 02 Jul 2015, **4**:180 (doi: 10.12688/f1000research.6743.1)
Latest published: 13 Jul 2015, **4**:180 (doi: 10.12688/f1000research.6743.2)

Abstract

Metagenomic sequence data can be used to detect the presence of infectious viruses and bacteria, but normal microbial flora make this process challenging. We re-analyzed metagenomic RNA sequence data collected during a recent outbreak of acute flaccid myelitis (AFM), caused in some cases by infection with enterovirus D68. We found that among the patients whose symptoms were previously attributed to enterovirus D68, one patient had clear evidence of infection with *Haemophilus influenzae*, and a second patient had a severe *Staphylococcus aureus* infection caused by a methicillin-resistant strain. Neither of these bacteria were identified in the original study. These observations may have relevance in cases that present with flaccid paralysis because bacterial infections, co-infections or post-infection immune responses may trigger pathogenic processes that may present as poliomyelitis-like syndromes and may mimic AFM. A separate finding was that large numbers of human sequences were present in each of the publicly released samples, although the original study reported that human sequences had been removed before deposition.

Open Peer Review

Referee Status: ✓

	Invited Referees 1
REVISED **version 2** published 13 Jul 2015	✓ report
version 1 published 02 Jul 2015	

1 **David Lipman**, National Institutes of Health USA

Discuss this article

Comments (3)

Corresponding author: Steven L. Salzberg (salzberg@jhu.edu)

How to cite this article: Breitwieser FP, Pardo CA and Salzberg SL. **Re-analysis of metagenomic sequences from acute flaccid myelitis patients reveals alternatives to enterovirus D68 infection [v2; ref status: approved 1, http://f1000r.es/5mz]** *F1000Research* 2015, **4**:180 (doi: 10.12688/f1000research.6743.2)

Grant information: This work was supported in part by the National Institutes of Health under grant R01-HG007196 and by the U. S. Army Research Office under grant number W911NF-14-1-0490.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare no competing interests.

First published: 02 Jul 2015, **4**:180 (doi: 10.12688/f1000research.6743.1)

REVISED **Amendments from Version 1**

We have revised our manuscript to make some small text amendments, in response to Charles Chiu's comments.

See referee reports

Background

Metagenomic shotgun sequencing, in which DNA or RNA is extracted from a tissue sample and then sequenced, has the potential to detect a wide range of infections. Deep whole-genome shotgun (WGS) sequencing can detect bacteria, viruses, and eukaryotic pathogens with equal effectiveness, as long as the infectious agent is similar to a species that has been previously sequenced. Sequencing databases already contain thousands of known species, and as this number grows, the sensitivity of WGS will grow as well.

In 2014, a large outbreak of infection with enterovirus D68 was associated with both severe respiratory illness and acute paralysis, which the U.S. Centers for Disease Control and Prevention (CDC) named acute flaccid myelitis (AFM)[1]. Samples collected from 48 patients were sequenced and shown to form a novel strain, Clade B1, based on phylogenetic analysis of 180 complete enterovirus D68 sequences[2]. The same study conducted metagenomic sequencing of cerebrospinal fluid (CSF) and/or nasopharyngeal (NP) swabs from 22 of these patients and found enterovirus D68 in some NP samples that were positive based on PCR testing.

The identification of species from a WGS sample is a challenging problem that has spurred the development of multiple new computational methods[3–5]. Because of the large size of next-generation sequencing data sets, these methods need to be very fast, but in the context of clinical diagnosis, they also need to be accurate. We downloaded the 31 next-generation sequencing (NGS) samples from the Greninger *et al.*[2] study (NCBI accession SRP055445) and re-analyzed them using a computational pipeline based on the recently developed Kraken metagenomic analysis software[4], a very fast and sensitive system that can be customized to use a database containing any species whose sequences are available.

Alternative infectious diagnoses in two subjects

Among the 22 subjects for which NGS data were available, we found at least two that had far greater numbers of sequences (reads) from a bacterial pathogen than from enterovirus D68. Neither subject had been reported in 2 as having a bacterial infection.

In one subject, US/CA/09-871, reported by Greninger *et al.*[2] as positive for enterovirus D68 through PCR and metagenomic NGS, we found in the NP swab sample an overwhelming presence of bacterial sequences from *Haemophilus influenzae*, a known cause of meningitis and neurological complications that was a common infection prior to the development of an effective vaccine.

Specifically, we identified 2,389,621 reads from *H. influenzae* in this subject, with the closest similarity to strain R2846. These reads comprise 93% of all microbial reads identified at the species level in the sample. Greninger *et al.*[2] reported 2,742 reads (in their Supplementary Table 4) matching enterovirus D68[2] but did not report finding any *H. influenzae* reads from this sample. Our analysis found 1,330 reads matching enterovirus D68.

To confirm the identity of these reads, we aligned them separately to the complete genome of *H. influenzae* R2846, and we found that the reads completely covered the genome. Dividing the genome into 100 kilobase windows, depth of coverage varied from 266–828 reads/100Kbp, with far deeper coverage as expected at the 16S ribosomal RNA genes.

The enterovirus D68 isolated from patient US/CA/09-871 differed from the others in that it appeared in 2009, well before the 2014 outbreak, and that it grouped with Clade C, phylogenetically distinct from Clade B1 that was associated with AFM. This patient was reported[2] as having respiratory illness but not AFM. The sequence evidence here suggests that the patient might have had complications from *H. influenzae*-associated infection, although no clinical or CSF data was available for our re-analysis.

In a second subject, US/CA/12-5837, we found a strikingly large number of reads from *Staphylococcus aureus* in the NP swabs. The two separate NGS files associated with this subject contained 6,858,453 and 1,343,806 reads, comprising 70% and 84% (respectively) of all non-human reads identified at the species level in each sample. The closest match was *S. aureus* subsp. *aureus* MRSA252, a methicillin-resistant strain. The coverage was deep enough, approximately 40X, that it would be possible to assemble this genome separately from the reads here (Figure 1).

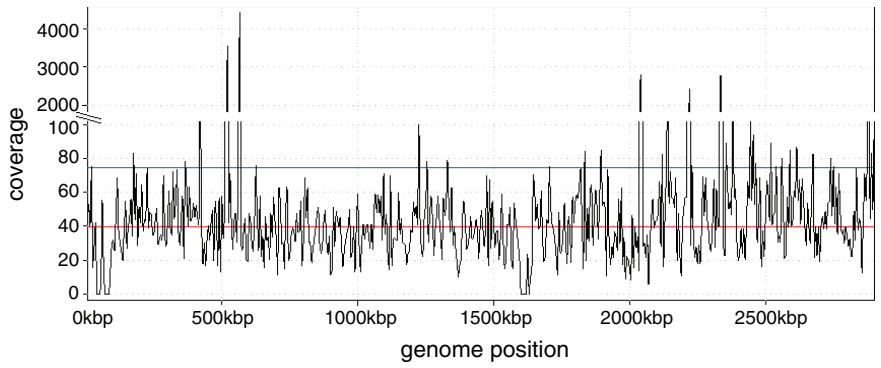

Figure 1. Depth of read coverage of the *S. aureus* MRSA252 genome using reads identified in the NGS sample from subject US/CA/12-5837. High peaks correspond to 16S rRNA genes. Red line: median coverage; blue line: mean coverage.

Greninger *et al.*[2] reported 2,790 reads from enterovirus D68 in this subject (our analysis found 1,641) but did not report any from *S. aureus*.

Patient US/CA/12-5837 was sampled in 2012, two years before the outbreak of AFM, although this patient was described in Greninger *et al.*[2] as positive for enterovirus D68 based on clinical PCR testing and metagenomic sequencing. This patient is reported to be one of the first patients with enterovirus-D68-positive AFM[2], but the sequence evidence indicates a severe *S. aureus* infection that might explain at least some of the patient's symptoms. *S. aureus* has been implicated in neurological complications such as myelitis[6] and meningitis[7] by mechanisms that involve not only direct invasion into the central nervous system (CNS), but also immunopathogenic responses triggered by superantigens that can target the CNS[8]. At a minimum, *S. aureus* infection was overlooked by the previous analysis. Although the potential role of bacterial infection in the neurological disease that affected these two subjects is difficult to assess because of the lack of clinical and CSF information, its involvement as a pathogenic co-factor should be evaluated.

Human reads included in database submission

The metagenomics data (NCBI accession SRP055445) released by Greninger *et al.*[2] comprise 43 files which cover 22 of the 48 subjects from their study (in their Supplementary Table 1); the study did not conduct NGS for all subjects. Our metagenomics pipeline identifies human reads at the same time that it searches for pathogens; therefore we scanned the data for human as well as microbial content. Greninger *et al.*[2] reported that all human sequences had been removed from these files. We found, however, that all samples contained large numbers of human reads, ranging from a low of 18,215 to a high of 6,159,868. These comprised as few as 0.5% to as many as 95.6% of the reads in each sample, as shown in Table 1.

The inclusion of human sequence data in the files deposited at NCBI was likely a result of a computational method (SURPI[5]) that was insufficiently sensitive. Although the exact cause cannot be determined here, it is well known that sequence alignment algorithms often trade speed for sensitivity; e.g., by allowing fewer mismatches, an aligner can process reads at a much higher rate, at the cost of missing some alignments. It is less clear why the very large numbers of matches to two bacteria were missed; for both these bacteria, complete genomes from multiple strains are available in GenBank. We used both the Kraken system[4] and the Bowtie2 aligner[9] to ensure both sensitivity and speed in our analysis.

Release of sequence data is highly valuable, if not essential, for reproducibility and validation of sequencing-based studies. Failure to filter human reads from a sample is not uncommon; a recent study[10] found that Human Microbiome Project samples, from which human DNA was supposed to have been removed, contain up to 95% human sequence. This suggests that future efforts to deposit microbiome data need to employ more sensitive computational screens in order to avoid the unintentional release of human sequence data.

Methods

Sequences were extracted from SRP055445 and each file was separately run through the Kraken program version 0.10.6-beta (https://github.com/DerrickWood/kraken)[4], which identifies species by

Table 1. Human reads found in metagenomic NGS samples from which human sequences were supposed to have been removed. Shown are the number of reads in each sample that clearly match the human genome and do not match any microbial species. AFM: acute flaccid myelits; NP: nasopharyngeal swap; CSF: cerebrospinal fluid.

Isolate	Run ID	Source	Number of human reads	%human
US/CA/12-5641	SRR1919640	NP	6,159,868	85.4
US/CA/12-5641	SRR1919641	NP	1,427,490	90.8
US/CA/12-5806	SRR1919642	NP	164,876	89.8
US/CA/12-5806	SRR1919643	CSF	202,677	95.5
US/CA/12-5807	SRR1919644	NP	160,719	94.1
US/CA/12-5807	SRR1919645	CSF	383,094	24.2
US/CA/12-5809	SRR1919646	NP	65,635	95.4
US/CA/12-5809	SRR1919647	NP	456,228	70.4
US/CA/12-5837	SRR1919648	NP	4,662,958	20.2
US/CA/12-5837	SRR1919649	NP	1,251,672	28.6
US/CA/14-5999	SRR1919650	CSF	3,046,664	89.9
US/CA/14-5999	SRR1919651	NP	1,407,842	71.0
US/CA/14-5999	SRR1919933	NP	174,140	68.5
US/CA/14-6000	SRR1919652	CSF	746,831	91.1
US/CA/14-6000	SRR1919653	NP	164,638	0.6
US/CA/14-6000	SRR1919934	NP	19,469	0.5
US/CA/14-6007	SRR1919654	CSF	352,391	85.4
US/CA/14-6010	SRR1919655	CSF	426,172	93.2
US/CA/14-6010	SRR1919656	NP	1,194,587	38.8
US/CA/14-6010	SRR1919935	NP	144,391	36.7
US/CA/14-6013	SRR1919657	NP	544,276	87.4
US/CA/14-6013	SRR1919658	NP	1,636,067	83.9
US/CA/14-6013	SRR1919936	NP	213,180	79.8
US/CA/14-6067	SRR1919659	CSF	567,263	3.9
US/CA/14-6067	SRR1919937	CSF	66,076	2.3
US/CA/14-6070	SRR1919660	CSF	578,579	4.3
US/CA/14-6070	SRR1919938	CSF	88,153	3.2
US/CA/14-6102	SRR1919661	CSF	791,143	82.4
US/CA/14-6102	SRR1919939	CSF	92,723	78.2
US/CO/13-60	SRR1919662	CSF	519,456	95.7
US/CO/13-60	SRR1919940	CSF	79,477	93.4
US/CO/14-86	SRR1919663	CSF	155,058	38.4
US/CO/14-86	SRR1919941	CSF	18,215	26.5
US/CO/14-88	SRR1919664	NP	453,411	3.8
US/CO/14-88	SRR1919942	CSF	39,899	2.7
US/CO/14-93	SRR1919665	CSF	758,650	96.6
US/CO/14-93	SRR1919943	CSF	123,250	95.3
US/CO/14-94	SRR1919666	NP	835,689	96.1
US/CO/14-94	SRR1919944	NP	131,998	95.2
US/CO/14-95	SRR1919667	CSF	352,679	2.8
US/CA/11-1767	SRR1919639	Culture	1,030,900	33.7
US/CA/10-786	SRR1919638	NP	130,044	0.5
US/CA/09-871	SRR1919637	CSF	384,285	11.0

comparison with a database of all 31-bp sequences in all species. The database included the human genome (version GRCh38.p2), all complete bacterial and viral genomes, selected fungal pathogens, and known laboratory vector sequences from the NCBI UniVec database (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec). Percentages of bacterial and viral reads in each sample were re-computed after excluding human and vector sequences. Reads matching more than one species were classified at the genus level or above. Reads from *H. influenzae* and *S. aureus* were re-aligned using Bowtie2 version 2.2.5[9], a very fast and sensitive program for alignment of NGS reads to a reference genome, with the --local option. Bowtie2 was also used to re-align all reads from US/CA/12-5837 and US/CA/09-871 to the sequence of multiple enterovirus D68 strains (GenBank accessions JX101846.1, AY426531.1, KM851231.1, KM892500.1, KM892501.1, KM881710.2, KP745751.1, KP745755.1, KP745757.1, KP745760.1, KP745764.1, KP745766.1, and KP745767.1). We report the highest number of reads matching any one of these strains.

Author contributions

SLS conceived the study. FBP ran the computational analyses. SLS, CAP, and FBP jointly analyzed the computational results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Grant information

This work was supported in part by the National Institutes of Health under grant R01-HG007196 and by the U. S. Army Research Office under grant number W911NF-14-1-0490.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Centers for Disease Control and Prevention. **Notes from the field: acute flaccid myelitis among persons aged ≤21 years - United States, August 1-November 13, 2014.** *MMWR Morb Mortal Wkly Rep.* 2015; **63**(53): 1243–1244.
PubMed Abstract
2. Greninger AL, Naccache SN, Messacar K, *et al.*: **A novel outbreak enterovirus D68 strain associated with acute flaccid myelitis cases in the USA (2012–14): a retrospective cohort study.** *Lancet Infect Dis.* 2015; **15**(6): 671–82.
PubMed Abstract | Publisher Full Text
3. Brady A, Salzberg SL: **Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models.** *Nat Methods.* 2009; **6**(9): 673–6.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biol.* 2014; **15**(3): R46.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Naccache SN, Federman S, Veeraraghavan N, *et al.*: **A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples.** *Genome Res.* 2014; **24**(7): 1180–92.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Saini M, Prasad K, Ling LM, *et al.*: **Transverse myelitis due to Staphylococcus aureus may occur without contiguous spread.** *Spinal Cord.* 2014; **52**(Suppl 2): S1–2.
PubMed Abstract | Publisher Full Text
7. Aguilar J, Urday-Cornejo V, Donabedian S, *et al.*: **Staphylococcus aureus meningitis: case series and literature review.** *Medicine (Baltimore).* 2010; **89**(2): 117–25.
PubMed Abstract | Publisher Full Text
8. Stach CS, Herrera A, Schlievert PM: **Staphylococcal superantigens interact with multiple host receptors to cause serious diseases.** *Immunol Res.* 2014; **59**(1–3): 177–81.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Ames SK, Gardner SN, Marti JM, *et al.*: **Using populations of human and microbial genomes for organism detection in metagenomes.** *Genome Res.* 2015.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status:

Version 2

Referee Report 15 July 2015

doi:10.5256/f1000research.7307.r9479

David Lipman
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

This paper provides a straightforward reanalysis of the metagenomic data presented in Greninger *et. al*. and does not contradict the basic interpretation of the results in that paper. As noted in the exchange of comments on this paper by Greninger *et al.* and Breitwieser *et. al.,* while Greninger *et al*. noted the staph and h.flu sequences in the supplementary data, they did not note these in the paper itself or in the table in the supplementary data that provided additional details - despite the fact that these were present in very high proportions. They also didn't report the presence of human sequence. If all these data were presented explicitly and in the main body of Greninger *et al*., a reader would be more aware of the challenges of these metagenomic approaches in infectious disease. I hope this paper will encourage this awareness and stimulate discussion so I note the comments on this paper from both sets of authors and I commend Breitwieser *et al.* for taking the time to respond as they have.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Discuss this Article

Version 2

Author Response (*Member of the F1000 Faculty*) 13 Jul 2015

Steven Salzberg, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, USA

We have made two small changes to correct two minor points raised by Chiu, Greninger, and Nacacche in their comments. First we reworded a sentence about sample US/CA/09-871 to clarify that the authors did not report any *Haemophilus influenzae* reads in this sample. (In Suppl. Table 3 they list a total count of bacterial reads found in US/CA/09-871, but in Suppl. Table 5, which lists bacteria by species name, no *H.*

influenzae reads are reported for this sample.) Second, we corrected the sentence where we said that patient US/CA/90-871 was reported as having encephalitis, which was incorrect. Greninger et al. list this patient as having a respiratory illness and this version of our paper now states that correctly.

Competing Interests: No competing interests were disclosed.

Version 1

Reader Comment (*Member of the F1000 Faculty*) 06 Jul 2015

Charles Chiu, Department of Medicine, University of California, San Francisco, USA

We thank Drs. Salzberg and Pardo for their response. Here are our replies addressing their new comments (in bulleted underline):

- "They don't disagree with our finding, but say that they already knew about it, pointing to their Supplementary Table 3. Suppl. Table 3 does indeed report large numbers of "bacterial reads" for these two samples, but it does not identify them further."

Yes, we knew about the detection of *Haemophilus influenzae* in sample US/CA/09-871 and MRSA (methicillin-resistant *Staphylococcus aureus)* in sample US/CA/12-5837, both of which are nasopharyngeal / oropharyngeal (NP/OP) swabs. We would be happy to provide the actual output from the SURPI pipeline (Naccache, et al., Genome Research, 2014) that shows the presence of these two bacteria, which were easily detected. As we previously stated, we reported the bacterial reads from the SURPI pipeline (5,614,487 reads from US/CA/09-871 and 28,676,383 reads from US/CA/12-5837); a "re-analysis" using a different algorithm was not necessary.

- "In contrast, Supplementary Table 5 in their paper includes specific read counts for 75 bacterial species that they found in most of their samples. These species include two species of Staphylococcus and one species of Haemophilus, but *S. aureus* and *H. influenzae* are not listed. Thus from the published paper, it is not possible to conclude that the authors were aware of either of these two species."

We respectfully request that the authors read our paper and tables more carefully. In Supplementary Table 5, we report the bacterial species found in only the cerebrospinal fluid (CSF) samples, not in nasopharyngeal / oropharyngeal (NP/OP) swabs. Again, we specifically state in our paper: ". As this protocol reduces sensitivity of detection and speciation for non-viral microbes (i.e. bacteria, fungi, and parasites), only viral sequences are shown for the NP/OP samples. "

- "Greninger, Naccache and Chiu also write that "the purpose of metagenomic NGS on NP/OP samples was to 'aid in the recovery of enterovirus D68 genome sequences and detect potential co-infections from other viruses." We agree that NGS can be valuable for this purpose, but we do not agree that findings of bacterial DNA - particularly when the bacteria dominate the sample, as in these two cases - should be ignored."

We did not "ignore" the findings of bacterial RNA/DNA. First, we emphasize that the sample library preparation on the NP/OP samples was a RNA, not a DNA preparation. It was post-treated with post-DNase to enrich for RNA viral sequences. As such, interpretation of bacterial sequence data from this library is problematic. It is entirely possible that other bacteria were present but their DNA was

selectively degraded, for instance. Also, such nuclease treatment will affect both the number and distribution of bacterial reads in the metagenomics sample, so attempting to derive quantitative information ("overwhelming presence of bacterial sequences") and, even worse, attempting to attribute clinical significance from the metagenomics data is not valid.

- Greninger et al. also make the point that they believe that "bacterial reads in NP/OP swabs from children most often reflect colonization/carriage and not infection." We agree: we would expect any metagenomics sample from the nasopharyngeal tract to contain many bacterial species. However, as we explain in our paper, the two samples in question have an overwhelming number of reads from just one species: in US/CA/09-871, over 93% of the reads were *H. influenzae* - 2.4 million reads - and in the two files for the other patient, 70% and 84% of the reads were *S. aureus* - over 8 million reads. It is possible that these merely represent colonization and were not causing disease. Even if so, we still feel it would have been important to acknowledge the presence of these pathogens in these samples (assuming the authors were aware of them) and to discuss each of them, at least briefly.

We re-iterate that it is dangerous to over-interpret the data as it appears as if Salzberg and colleagues are doing here. Please see the previous paragraph on why the number and distribution of bacterial metagenomic reads in these cases are not reliable. As mentioned before, we do report the number of bacterial reads found in NP/OP samples in Supplementary Table 3. We did not discuss them given that (1) the focus for this study was on looking for clinical associations with AFM, (2) analysis of the NP/OP samples were problematic because the preparation was not suitable / reliable for bacterial metagenomics analysis and the difficulty in discriminating colonization from genuine infection, and (3) the presence of reads to many and multiple bacterial species across all of the NP/OP samples, not just the.two containing MRSA or *Haemphilus influenza*.

Our focus was also on examination of "sterile sites" (e.g. cerebrospinal fluid and blood) for sequences to pathogens, since it is much easier to demonstrate that an infectious agent is associated or even causal if it is detected in a sterile site. We did not think that specifically mentioning *Haemphilus influenza*e from a positive control NP/OP sample would be relevant to the study, nor MRSA from the nose/throat of an AFM patient in part because it would be very difficult to distinguish colonization from infection. Also note that infectious disease diagnosis of MRSA infection is primarily made by positive cultures or detection of the bacteria from sterile or invasive sites such as blood, cerebrospinal fluid, and bronchoalveolar lavage fluid and not from nasal swabs which are mainly used to screen for MRSA colonization.

- "Greninger et al. express concern that our analysis suggests that MRSA or *H. influenzae* infections are involved in the pathogenesis of AFM. We do not suggest this, but we suggest that the role of bacterial infection as a co-factor should be evaluated in patients suspected of having AFM. Although not the point of our study or of the Greninger et al. study, this is an important issue in clinical practice and for future studies of patients suspected to have AFM, due to the potential role of bacteria in triggering inflammatory myelopathies, encephalomyelitis syndromes such as Acute Disseminated Encephalomyelitis (ADEM), or superantigen-induced disorders, which may present clinically as longitudinal extensive myelopathies that may resemble AFM."

We agree that bacterial infection can be a co-factor. However, as previously mentioned, the library preparation from NP/OP was not suitable for bacterial metagenomics analysis. To evaluate this properly, the samples can and should be prepared as bacterial metagenomics libraries, and a complete analysis of the results, not just mentioning the *Haemophilus influenza*e and MRSA – as done here -- but also all of the other bacteria in the other samples, performed. Even after this analysis is done, much more study would need to be done to investigate the role of bacterial infection as a co-factor. The inadequate

sample preparation for bacterial metagenomics analysis and non-sterile sample type (NP/OP) make our NP/OP data not suitable for investigating "the role of bacterial infection as a co-factor".

- The authors do point out one error in our paper, where we stated that patient US/CA/09-871 was reported (by them) as having "encephalitis and severe respiratory illness." This was an error on our part, and we will submit a revised manuscript where we correct this to read "severe respiratory illness."

We feel that this is a major error in the paper, as one of the two patients did not even have acute flaccid myelitis. Also, it is incorrect to say "severe respiratory illness"; this was a patient with an upper respiratory infection that may not have been clinically severe. We feel that it is inappropriate in general to report the results of metagenomics analyses in this fashion without understanding the clinical context and how the sample was collected and prepared. Note also that given this error, the title becomes even more misleading: "re-analysis of metagenomics sequences from acute flaccid myelitis patients reveals alternatives to enterovirus D68 infection."

- "Regarding our second main criticism - that the authors inadvertently left large numbers of human reads in their data, which were supposed to have been filtered to remove human DNA - the authors agree. As we mentioned in our paper, this problem has occurred elsewhere too: J. Allen and colleagues recently showed (Ames et al 2015, cited in our paper) that some of the Human Microbiome Project samples, all of which were supposed to have been cleaned of human reads, contain up to 95% human reads. We believe it is vitally important that the biomedical community be aware that the choice of computational methods is a very critical one, particularly when analyzing the large data sets that are becoming ever more common."

We agree and have taken steps to ensure patient privacy (requesting that SRA data be removed from the database). However, we also believe that it is unfair to target our paper specifically, as many other earlier published studies including the Human Microbiome Project, as the authors point out, have the same issues. We will be erforming a bioinformatics analysis comparing the different computational methods for their ability to remove human reads.

- We believe it is vitally important that the biomedical community be aware that the choice of computational methods is a very critical one, particularly when analyzing the large data sets that are becoming ever more common.

Although we agree with this statement, we believe that a "re-analysis" was not necessary for our paper as we detected both bacteria described by Salzberg and colleagues. We would have been happy to provide this data upon request. In addition, the re-analysis has numerous errors and issues of over-interpretation that did not take into account the clinical context and not understanding important details related to the study (e.g. upper respiratory infection does not imply "severe respiratory illness"; the NP/OP sample preparation was sub-optimal and problematic for bacterial metagenomics analysis; our patient with *Haemphilus influenza*e did not have acute flaccid myelitis). We believe that the re-analysis by Salzberg and colleagues simply confirms the results from our published analysis, although there are obvious differences in clinical interpretation.

Alexander Greninger, MD/PhD
Samia Nacacche, PhD
Charles Chiu, MD/PhD

Competing Interests: We are authors from the paper by Greninger, et al., (2015) published in Lancet Infectious Diseases.

Author Response (*Member of the F1000 Faculty*) 03 Jul 2015

Steven Salzberg, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, USA

We thank Drs. Greninger, Naccache, and Chiu for their response, which makes some valid points that we will comment on further here. But first we wish to acknowledge that our re-analysis does not affect the main finding of their paper; i.e., that the 2014 outbreak of enterovirus D68 represented a novel strain, as shown by their phylogenetic analysis. We don't question that finding, and we think it is an important contribution to the understanding of acute flaccid myelitis.

In their comment, they respond to our first main criticism that they appear to have missed the fact that two of their samples were dominated by the presence of *Haemophilus influenza* (sample US/CA/09-871) and *Staphylococcus aureus* (sample US/CA/12-5837). They don't disagree with our finding, but say that they already knew about it, pointing to their Supplementary Table 3. Suppl. Table 3 does indeed report large numbers of "bacterial reads" for these two samples, but it does not identify them further.

In contrast, Supplementary Table 5 in their paper includes specific read counts for 75 bacterial species that they found in most of their samples. These species include two species of Staphylococcus and one species of Haemophilus, but *S. aureus* and *H. influenzae* are not listed. Thus from the published paper, it is not possible to conclude that the authors were aware of either of these two species.

Greninger, Naccache and Chiu also write that "the purpose of metagenomic NGS on NP/OP samples was to 'aid in the recovery of enterovirus D68 genome sequences and detect potential co-infections from other viruses." We agree that NGS can be valuable for this purpose, but we do not agree that findings of bacterial DNA - particularly when the bacteria dominate the sample, as in these two cases - should be ignored.

Greninger et al. also make the point that they believe that "bacterial reads in NP/OP swabs from children most often reflect colonization/carriage and not infection." We agree: we would expect any metagenomics sample from the nasopharyngeal tract to contain many bacterial species. However, as we explain in our paper, the two samples in question have an overwhelming number of reads from just one species: in US/CA/09-871, over 93% of the reads were *H. influenzae* - 2.4 million reads - and in the two files for the other patient, 70% and 84% of the reads were *S. aureus* - over 8 million reads. It is possible that these merely represent colonization and were not causing disease. Even if so, we still feel it would have been important to acknowledge the presence of these pathogens in these samples (assuming the authors were aware of them) and to discuss each of them, at least briefly.

Greninger et al. express concern that our analysis suggests that MRSA or *H. influenzae* infections are involved in the pathogenesis of AFM. We do not suggest this, but we suggest that the role of bacterial infection as a co-factor should be evaluated in patients suspected of having AFM. Although not the point of our study or of the Greninger et al. study, this is an important issue in clinical practice and for future studies of patients suspected to have AFM, due to the potential role of bacteria in triggering inflammatory myelopathies, encephalomyelitis syndromes such as Acute Disseminated Encephalomyelitis (ADEM), or

superantigen-induced disorders, which may present clinically as longitudinal extensive myelopathies that may resemble AFM.

The authors do point out one error in our paper, where we stated that patient US/CA/09-871 was reported (by them) as having “encephalitis and severe respiratory illness.” This was an error on our part, and we will submit a revised manuscript where we correct this to read “severe respiratory illness.”

Regarding our second main criticism - that the authors inadvertently left large numbers of human reads in their data, which were supposed to have been filtered to remove human DNA - the authors agree. As we mentioned in our paper, this problem has occurred elsewhere too: J. Allen and colleagues recently showed (Ames et al 2015, cited in our paper) that some of the Human Microbiome Project samples, all of which were supposed to have been cleaned of human reads, contain up to 95% human reads. We believe it is vitally important that the biomedical community be aware that the choice of computational methods is a very critical one, particularly when analyzing the large data sets that are becoming ever more common.

Steven Salzberg, Ph.D.
Carlos Pardo, M.D.

Competing Interests: We are authors of the F1000 Research paper being commented upon here.

Reader Comment (*Member of the F1000 Faculty*) 03 Jul 2015

Charles Chiu, Department of Medicine, University of California, San Francisco, USA

This manuscript raises two main criticisms of our paper in their re-analysis. Here we directly address the 2 main points:

1. The authors claim that bacterial reads were seen in the nasopharyngeal / oropharygneal swab (NP/OP) metagenomic data that were "missed" in the original study. They were able to assemble two genomes, from *Haemophilus influenzae* and methicillin-resistant *Staphylococcus aureus*. We would like to highlight that these reads and bacteria were not missed but instead we did not discuss the bacterial/fungal portion of the NP/OP swabs in the manuscript due to difficulties in clinical interpretation.

As quoted from the supplementary section of our manuscript published in *Lancet Infectious Diseases*, with our emphasis underlined:

- **Lancet ID Supplementary Material:** NGS libraries constructed from NP/OP samples were treated with DNase following nucleic acid extraction to reduce background from the human host and bacterial flora. <u>As this protocol reduces sensitivity of detection and speciation for non-viral microbes (i.e. bacteria, fungi, and parasites), only viral sequences are shown for the NP/OP samples</u>. The ability to detect DNA viruses is also impacted by the use of DNase and we cannot exclude the possibility that our data is biased by reduced sensitivity for detection of DNA viruses.

- The authors also state that we "did not report finding any bacterial reads from this sample" but that does not mean that we did not detect any bacterial reads. In fact, we detected both the *Haemophilus influenzae* and methicillin-resistant *Staphylococcus aureus* reported by the authors:
  - The number of bacterial reads found in all metagenomic samples analyzed is shown in Supplementary Table 3, Column N. Bacterial reads for the two samples in question are listed as US/CA/09-871: 5,614,487 bacterial reads and US/CA/12-5837: 28,676,383 bacterial reads.
  - In the main text we also state that the purpose of metagenomic NGS on NP/OP samples was to "to aid in the recovery of enterovirus D68 genome sequences and detect potential co-infections from other viruses".

- In addition, although we reported the presence of bacterial reads in the NP/OP data, detected using SURPI/SNAP, we did not discuss this in the paper due to a number of reasons:
  - our focus on looking for CSF (bacterial viral fungal and parasite) pathogens in the setting of AFM
  - our clinical perspective that the presence of bacterial reads in NP/OP swabs from children most often reflect colonization / carriage and not infection (see http://www.biomedcentral.com/1471-2180/10/59, http://www.ncbi.nlm.nih.gov/pubmed/15999003, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3962756/, http://www.ncbi.nlm.nih.gov/pubmed/12394812, etc. for papers on nasal carriage of *Staphylococcus* and *Haemophilus* in healthy children)
  - our treatment of the nucleic acid extracts with DNase to help reduce host background, which would bias accurate metagenomic interpretation and the accurate counting of bacterial and fungal reads since this procedure degrades bacterial / fungal genomic DNA
  - The NP/OP microbiome in healthy individuals is dominated by bacteria, including potential pathogens, and we chose not to comment on this as it is well described and not the focus of our paper. The interpretation of metagenomic data without a clinical understanding of infectious diseases and the microbiology of specific bacterial species can lead to incorrect and even harmful conclusions – we chose to interpret and report on our data in the context of our knowledge of infectious diseases and clinical microbiology.

- Of note, the sample in question with *Haemophilus influenzae* (US/CA/09-871) came from a 2009 case of upper respiratory infection alone. This subject did not have either encephalitis or acute flaccid myelitis (Table 2 and results in Greninger et al.). The statements in the manuscript incorrectly state that "This patient was reported as having encephalitis and severe respiratory illness..." and "the sequence evidence here suggests that the patient might have had complications from *H. influenzae*-associated encephalitis or encephalomyelitis...". These statements as well as the title of the manuscript are therefore incorrect.

1. The authors point out that there are residual human reads in the deposited data.

We acknowledge that we have been using SNAP, which is a global aligner, to extract out human reads for our SRA submissions. This algorithm will miss human reads because of low-complexity sequences at the ends, residual adapters, etc. We agree that we probably should have used a local aligner such as BLASTn at a low threshold level to more completely extract out human reads, or a k-mer approach such as Kraken. We appreciate the authors' point on the importance of clearing human sequences from metagenomic data if we are to completely deidentify the sample. We do not know whether any of the residual human reads are potentially identifying, so for now will delete the data from SRA.

The re-analysis presented here gives the erroneous impression that bacterial colonization has a strong association with the devastating acute flaccid myelitis (AFM) syndrome seen in our patients. As we note above, *Haemophilus influenzae* was found in a control patient with no neurological symptoms, and is thus not relevant to AFM. While MRSA colonization may possibly be a factor in AFM, it is extremely unlikely given its detection is a single case and the fact that MRSA nasal carriage in healthy children is well-described. We should also point out that bacterial cultures of cerebrospinal fluid (CSF) from all of the patients in the study were negative and that metagenomic next-generation sequencing did not reveal any evidence of bacterial infection in the central nervous system (CNS).

We truly appreciate the discussion demonstrated towards understanding this important and serious clinical syndrome and for providing an opportunity for us to respond by using the F1000 platform. We would appreciate it in the future if the authors would have the courtesy of calling or sending us an e-mail. Our group is highly collaborative and we would have been happy to discuss the results of our study and our interpretation in detail.

We also thank the authors for their interest in our paper, its publicly available data, and in highlighting the importance of metagenomic sequence analysis for clinical diagnostics. We look forward to productive discussion in the future on the many clinical applications of this genomic technology that will greatly benefit patients in the future.

Alexander Greninger, M.D./Ph.D.
Samia Naccache, Ph.D.
Charles Chiu, M.D./Ph.D.

Competing Interests: We are authors on the original paper referenced by this manuscript.